AI-Driven Text Analysis in the Political Economy of Sustainability

Hybrid Retrieval-Augmented Generation and LLM Multi-Agent Approach

Bastián González-Bustamante
Leiden University
✉ b.a.gonzalezbustamante@fgga.leidenuniv.nl

Natascha van der Zwan
Leiden University
✉ n.a.j.van.der.zwan@fgga.leidenuniv.nl

Presentation at the ECPR Joint Sessions of Workshops, Charles University
Prague, Czech Republic, May 20-23, 2025

Introduction

Challenges in Situating Sustainable Finance

Conceptual Challenge

How to place sustainable finance within existing politico-economic typologies

Empirical Challenge

Lack of official SI data and consensus on sustainable investment definition

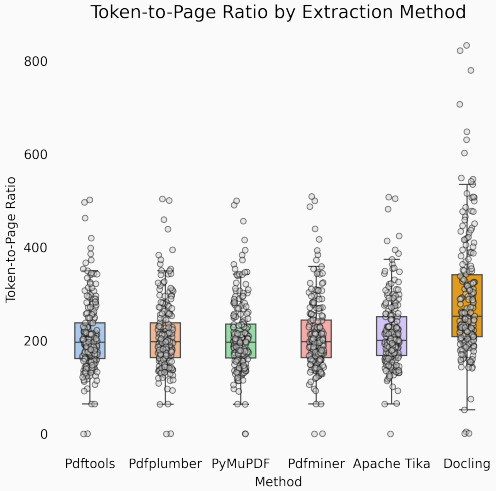
Greenwashing Concern

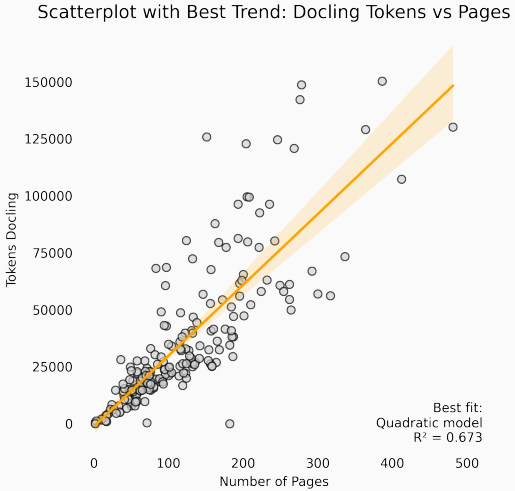
Difficulty in distinguishing genuine sustainable practices from greenwashing

Artwork by DALL·E model

Research Design

Benchmark of Annual Reports Parsing Methods





Note. Two observations were excluded from the plots because of their high number of tokens.

Data Overview

32

Countries

Top asset owners from these countries for the moment

24K

Pages

Only annual reports, but we are including specific reports and disclosures

7.6M

Tokens

Tokens for processing and analysis for the moment

Model Selection Benchmark

96 LLMs

CAP's major policy topics environment and energy in English for

(1) **error rate analysis**
(2) **orchestration model selection**

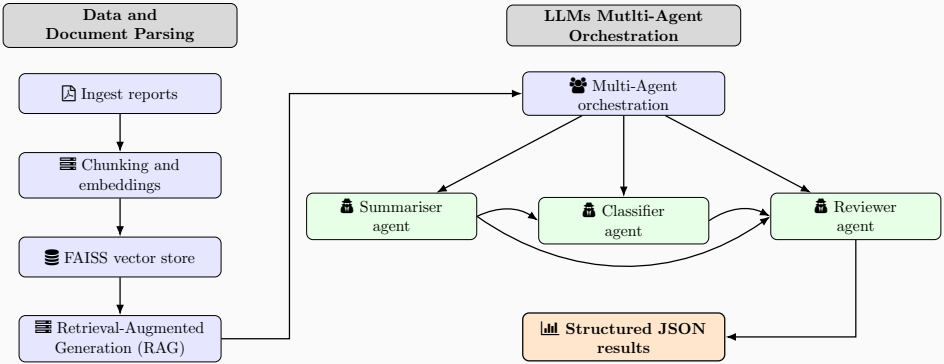
SOTA closed-source LLMs

o3, o4-mini, GPT-4.1, GPT-4.5-preview, Gemini 2.5, Grok 3 Beta, Claude 3.7 Sonnet

SOTA open-source LLMs

Llama 4 Maverick (400B) and Scout (107B), Mistral 3.1 (24B), Llama 3.3 (70B), DeepSeek-R1 (671B), DeepSeek-V3 (671B)

RAG and Multi-Agent Orchestration Pipeline



Multi-Agent Orchestration

Summariser Prompt

You are an expert financial analyst specialising in sustainable finance disclosures. Using only the retrieved context, write a concise (200-300 word) neutral summary. Focus on these six analytical frames: emissions reporting, climate leadership, sustainable investments, green finance, climate-related risks, energy transition.

After the prose summary. Estimate how prevalent each frame is as a percentage of total sentences in the underlying document (0-100, rounded). Output a single JSON object named frame_prevalence, with exactly these six keys, each mapped to its % value (integer). Nothing else may follow that JSON.

Multi-Agent Orchestration

Classifier Prompt

You are an ESG taxonomy specialist. Determine how prominently the document frames every topic below, using both the prose summary and the frame_prevalence JSON (sentence-share %): emissions reporting, climate leadership, sustainable investments, green finance, climate-related risks, energy transition.

Label each frame with one of: Not Mentioned ($< 1\%$), Peripheral ($\sim$1-5%), Moderate ($\sim$6-15%), Central ($> 15\%$).

Respond only with JSON having exactly these six keys: emissions_reporting, climate_leadership, sustainable_investments, green_finance, climate_related_risks, energy_transition.

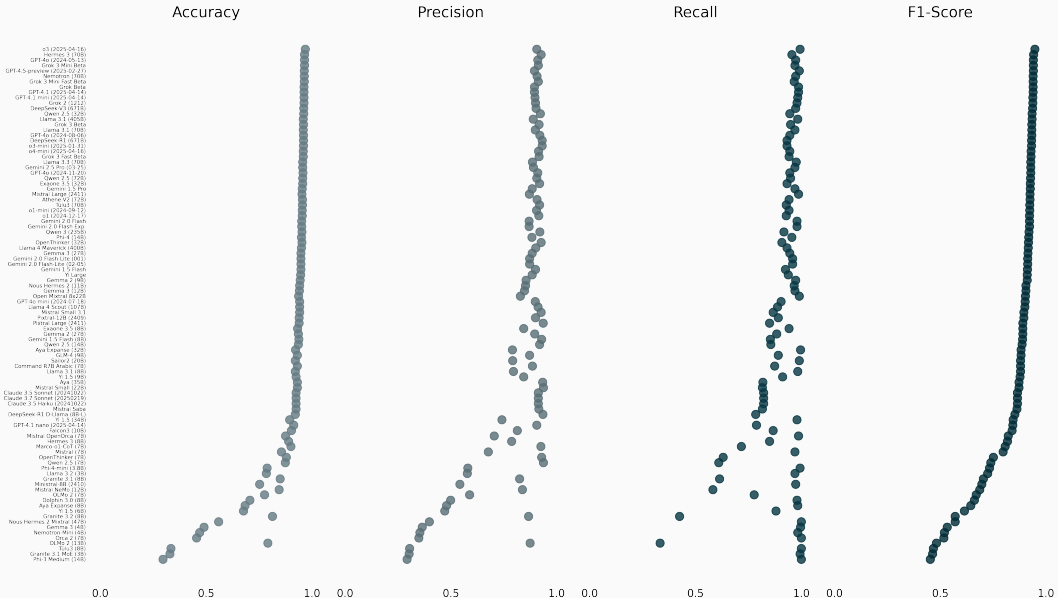
Multi-Agent Orchestration

Reviewer Prompt

You are the senior reviewer. Evaluate the summary, the frame_prevalence JSON, and the classification below. If anything is missing, mislabelled, inconsistent, or exceeds the 300-word limit, produce corrected versions.

Return only a JSON object with three keys: revised_summary (a refined summary ($\leq$ 300 words), revised_frame_prevalence (the prevalence JSON in the same format), revised_classification (the six-key classification JSON).

Preliminary Results


Accuracy
Precision
Recall
F1-Score
o3 (2025-04-16)
Hermes 3 (70B)
GPT-4.1 Mini (2025-04-14)
Grok 3 Mini Beta
GPT-4.5-preview (2025-02-27)
Nemotron (70B)
Grok 3 Mini
Grok 3 Beta
GPT-4.1 (2025-04-14)
Grok 2 (1212)
DeepSeek V3 (0324)
Qwen 2.5 (32B)
Llama 3.1 (405B)
Grok 3 Beta
Llama 3 (70B)
GPT-4o (2024-08-06)
DeepSeek R1 (0116)
o3-mini (2025-01-31)
o4-mini (2025-04-16)
Grok 3 Fast Beta
Llama 3.1 (70B)
Gemini 2 V-Pro (03-25)
GPT-4o (2024-11-20)
Qwen 2.5 (72B)
Cogvent 3.5 (32B)
Gemini 1.5 Pro
Mistral Large (2411)
Athene V2 (72B)
Grok1 (70B)
o1 mini (2024-09-12)
o1 (2024-12-17)
Llama 4 Maverick
Gemini 2.0 Flash Exp
Qwen 3 (235B)
Qwen 2.5 (14B)
Open Mistral Nemo
Llama 3 Instruct (70B)
Gemma 3 (27B)
Gemini 2.0 Flash Lite (001)
Gemini 2.0 Flash Lite (Cogvent)
Gemini 1.5 Flash
Yi Large
Gemma 2 (9B)
Nous Hermes 2 (11B)
Qwen 2.5 (7B)
Open Mistral 8x22B
GPT 4o mini (2024-07-18)
Llama 4 Scout (109B)
Qwen 2.5 (32B)
Mistral 7.3B (2409)
Mistral Large (2411)
Cogvent 3.5 (8B)
Qwen 2.5 (3B)
Gemini 1.5 Flash (8B)
Qwen 2.5 (1.5B)
Aya Expanse (32B)
GLM-4 (9B)
Qwen 2.5 (0.5B)
Command R7B Arabic (7B)
Llama 3.1 (8B)
Yi 1.5 (9B)
Qwen 2.5 (14B)
Mistral Small (2501)
Claude 3.5 Sonnet (20241022)
Claude 3.7 Sonnet (20250219)
Claude 3.7 Haiku
Mistral Tuzu
DeepSeek R1 Distill Llama (8B)
Yi 1.5 (34B)
GPT-4.1 nano (2025-04-14)
Qwen 3 (8B)
Mistral OpenOrca (7B)
Hermes 3 (8B)
Mistral 7.3B (7B)
Yi 1.5 (6B)
OpenThinker (7B)
Qwen 2.5 (7B)
Phi-4-mini (3.8B)
Qwen 3 (14B)
Granite 3.1 (8B)
Ministral 8B (2410)
Mistral NeMo (12B)
OLMo 2 (7B)
Dolphin 3 (8B)
Aya Expanse (8B)
Yi 1.5 (9B)
Granite 3.3 (8B)
Nous Hermes 2 Mixtral
Gemma (14B)
Nemotron Mini (4B)
Orca 2 (7B)
OLMo 2 (13B)
Qwen (4B)
Granite 3.1 MoE (3B)
Phi-3 Medium (14B)
0.0
0.5
1.0

Per-Frame Metrics (35% of the corpus)

Frame	Agent	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score
Emissions reporting	Classifier	0.986	0.986	0.986	0.985
Climate leadership	Classifier	0.929	0.980	0.929	0.943
Emissions reporting	Summariser	0.929	0.961	0.929	0.939
Climate leadership	Summariser	0.900	0.943	0.900	0.915
Energy transition	Classifier	0.843	0.841	0.843	0.822
Green finance	Classifier	0.800	0.809	0.800	0.797
Sustainable investments	Summariser	0.800	0.886	0.800	0.793
Energy transition	Summariser	0.786	0.809	0.786	0.792
Green finance	Summariser	0.771	0.793	0.771	0.766
Climate-related risks	Summariser	0.757	0.819	0.757	0.744
Climate-related risks	Classifier	0.686	0.749	0.686	0.655
Sustainable investments	Classifier	0.629	0.820	0.629	0.648

Note. Reviewer agent as ground-truth.

Takeaways and Next Steps

Takeaways and Next Steps

☑ **Goof Performance.** The summariser reaches 0.824 accuracy and 0.826 weighted F1-score against the reviewer yardstick. The classifier follows a fraction behind, leaving a residual gap of 1.2 points.

☑ **Hybrid Design.** RAG backbone, orchestration and mixed licensing strategy. Swapping in components is possible.

☑ **Next Steps.** There are a number of immediate improvements and next steps:

- **Some Problems.** (1) Most of that shortfall is concentrated in frames as sustainable investments and green finance. (2) Closed-source summariser and reproducibility.
- **Improvements.** (1) Classifier prompt needs to tie salience labels more tightly. (2) Temperature and nucleus-samples sweeps can probe whether a touch of controlled randomness. (3) Human-in-the-loop.

Thank you very much!

Do you have any questions?

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com

Presentation compiled with LaTeX and some

Download the latest version from GitHub